第 39 卷第 8 期
录用定稿

计算机应用研究
Application Research of Computers

Vol. 39 No. 8
Accepted Paper

基于非鲁棒特征的图卷积神经网络对抗训练方法

承　琪，朱洪亮†，辛　阳

(北京邮电大学 网络空间安全学院，北京 100876)

摘　要：图卷积神经网络可以通过图卷积提取图数据的有效信息，但容易受到对抗攻击的影响导致模型性能下降。对抗训练能够用于提升神经网络鲁棒性，但由于图的结构及节点特征通常是离散的，无法直接基于梯度构造对抗扰动，而在模型的嵌入空间中提取图数据的特征作为对抗训练的样本，能够降低构造复杂度。借鉴集成学习思想，提出一种基于非鲁棒特征的图卷积神经网络对抗训练方法 VDERG，分别针对拓扑结构和节点属性两类特征，构建两个图卷积神经网络子模型，通过嵌入空间提取非鲁棒特征，并基于非鲁棒特征完成对抗训练，最后集成两个子模型输出的嵌入向量作为模型节点表示。实验结果表明，提出的对抗训练方法在干净数据上的准确率平均提升了 0.8%，在对抗攻击下最多提升了 6.91%的准确率。

关键词：图卷积神经网络；集成学习；非鲁棒特征；对抗训练

中图分类号：TP183　　doi: 10.19734/j.issn.1001-3695.2022.01.0012

Graph neural networks adversarial training with non-robust features

Cheng Qi, Zhu Hongliang†, Xin Yang

(*Cyber Security*, *Beijing University of Posts & Telecommunications*, *Beijing* 100876, *China*)

Abstract: Graph convolutional neural networks can distill the effective information of graph data through graph convolution. However, the graph convolutional neural network show vulnerability to adversarial attack, which leads to the degradation of model performance. Adversarial training can be used to improve the robustness of neural networks. However, since the structure and node features of graphs are usually discrete, it is impossible to directly construct adversarial examples based on gradients. Therefore, distilling feature of graph data in the embedding space of models as adversarial examples can reduce the complexity of adversarial training. By using the idea of the idea of ensemble learning, this paper innovatively proposes an adversarial training method based on non-robust features distillation for graph convolution network, VDERG. The method constructed two graph convolution neural networks as sub models from the two types of features of topology and node attributes respectively. Sub models distilled non-robust features through embedding space and used these features to implement adversarial training. Finally, the method combined the embedding given by the two sub models as the nodes' vectors. Experimental results show that the adversarial training strategy improves the accuracy of graph convolution neural networks in clean data by 0.8% on average, and improves the accuracy by 6.91% at most under adversarial attack.

Key words: graph convolutional neural network; ensemble learning; non-robust features; adversarial training

0　引言

图作为一种具有普遍性的数据结构，可以广泛用于表示不同领域中的系统，例如经济领域(交易网络)、社会科学领域(社交网络和引文网络)、自然科学领域(分子结构)和知识图等。近年来图神经网络(graph neural network, GNN)在学习图表示方面取得了令人瞩目的成果。其中，图卷积神经网络(graph convolutional neural network, GCN)通过利用边的信息对节点信息进行聚合生成节点表示，在图信息的提取方面效果显著。从图中提取出的特征可以用于节点分类、链路预测、图分类等任务，在数据挖掘、推荐系统等领域有着广泛的应用。

已有研究证明缺乏鲁棒性的神经网络容易受到对抗攻击的影响，即加入了微小扰动的对抗样本，会大大降低神经网络的模型表现[1]。Dai 等人[2]发现随机丢弃节点间的边就能对图神经网络造成较好的攻击效果。GCN 的脆弱性可能在其应用领域导致安全问题，例如在信用检测系统中，欺诈者可以通过与几个高信用用户伪装多个交易，从而在模型检测中获得“高信用用户”的虚假结果[2]。因此开始有大量研究针对提升 GCN 鲁棒性展开。

对抗训练[3]被广泛用于提升神经网络的鲁棒性：通过在模型训练过程中加入对抗样本，使神经网络适应对抗扰动，从而提升对抗攻击下的模型表现。现有针对 GCN 的对抗训练方法研究主要集中于针对单个模型构造扰动正则项或修改图结构，少有研究从集成的角度借助多个分类器的学习能力提升模型的鲁棒性。

对抗攻击的特征之一是在神经网络间具有泛化性，而通过集成多个神经网络模型分别进行对抗训练，能够使总体模型学习到更全面的特征信息，从而提升模型鲁棒性。文献[4,5]指出，基于集成学习的防御算法效果依赖于子模型的多样化。只有使子模型分别学习到不同的特征，才能避免对抗扰动在子模型间迁移，有效提升总体模型的防御能力。考虑到图数据的特征，Wu 等人[6]构造了包含两个子模型的集成模型，分别针对拓扑结构信息和节点属性信息进行模型训练，以此提升 GNN 鲁棒性。但仅仅通过在结构信息和属性信息上分开训练子模型，没有考虑对抗攻击的攻击特点，在结构信息和属性信息都受到攻击的情况下仍可能产生较大的预测偏差。

收稿日期：2022-01-09；**修回日期：**2022-03-11

作者简介：承琪(1997-)，女，福建厦门人，硕士研究生，主要研究方向为网络安全；朱洪亮(1982-)，男(通信作者)，河南漯河人，副教授，博士，主要研究方向为大数据安全、网络行为分析、网络安全(zhuhongliang@bupt.edu.cn)；辛阳(1977-)，男，山东海阳人，教授，博士，主要研究方向为网络安全与智能信息处理、网络空间安全、人工智能、存储灾备等.

对神经网络进行模型训练本质上是从图数据中学习特征的过程，而学习到的有利于提升模型表现的特征会对对抗扰动表现出不一样的敏感度，基于不同的对抗扰动敏感度，可以将这些特征分为鲁棒特征和非鲁棒特征两类。数据中的鲁棒特征即使在对抗攻击下，仍能保持稳定性，帮助模型学习正确的有效信息，而非鲁棒特征则会被对抗扰动窜改，使模型在训练过程中学习错误的信息，进而导致模型表现降低。模型学习到的非鲁棒特征导致了模型的脆弱性，但目前基于非鲁棒特征进行对抗训练的研究都针对图像数据展开，鲜有研究利用图数据上的非鲁棒特征提升模型鲁棒性。

基于上述问题，本文旨在探索图数据中的非鲁棒特征对于提升 GCN 模型鲁棒性的作用，并结合图数据中的结构信息和节点属性信息，给出非鲁棒特征定义及提取方法。并且，基于 GCN 从图数据中学习到的非鲁棒特征以及集成学习方法，提出一套基于非鲁棒特征的鲁棒图卷积神经网络集成模型(vulnerabilities distillation of ensembles for robust graph neural networks, VDERG)。VDERG 利用图卷积层后的嵌入向量，分别从结构信息和属性信息中提取非鲁棒特征，并以此对两个子模型分别进行对抗训练，使两个子模型分别适应节点关系和节点属性上的对抗扰动，然后集成两个子模型的节点嵌入向量，输入映射函数作为最终预测结果。实验证明本文提出的防御算法能够有效提高图卷积神经网络模型的鲁棒性。

本文的主要贡献如下：a)定义了图数据上的非鲁棒特征，并结合图数据特点，从结构信息和属性信息两方面给出非鲁棒特征提取方法；b)提出一种基于集成学习的鲁棒性图卷积神经网络算法，通过非鲁棒特征对子模型进行对抗训练，并使不同子模型分别从结构信息和属性信息中学习图信息，集成节点向量表征，有效抵御对抗攻击影响。

1　相关工作

1.1　对抗训练

图卷积神经网络可以看做卷积神经网络在图数据上的迁移变种，由于二者相似的卷积机制，图卷积神经网络同样容易受到对抗攻击干扰。对于图 $\mathcal{G}=(A,X)$ 上节点水平的对抗攻击，攻击者的目标是找到一个图结构 $\hat{\mathcal{G}}=(\hat{A},\hat{X})$，能够最大化目标节点 V_t 在图神经网络模型 f_θ 上的损失值 $\mathcal{L}_{atk}(f_\theta(\hat{\mathcal{G}}))$，其中需要约束对抗扰动不易被察觉，即 $\|\hat{A}-A\|_0+\|\hat{X}-X\|_0\leqslant\Delta$。由于图数据上的任务特性，目前大多数的对抗攻击为投毒攻击，即攻击者在训练数据集中加入对抗样本实施攻击[7]。

近年来由于图卷积神经网络在节点表征上的强大表达能力，有许多研究开始关注如何提升 GCN 模型的鲁棒性。对抗训练已经在提升卷积神经网络等模型的鲁棒性上取得了显著成果，也被众多学者借鉴用于提升 GCN 模型的鲁棒性。对抗训练在模型训练过程中生成对抗样本，并同时最小化模型在对抗样本上的损失值，即 $\min_\theta \mathcal{L}_{train}(f_\theta(\hat{\mathcal{G}}))$。Dai 等人[2]通过在模型训练过程中随机丢弃图中的边从而对扰动邻接矩阵，但这种训练方法只降低了 1%的攻击成功率。Dai 等人[8]基于 DeepWalk[9]，提出在嵌入空间中加入噪声以进行针对投毒攻击情境下的对抗训练，提升了 DeepWalk 在节点分类任务上的泛化能力。这种对抗训练方法可以扩展到一系列节点嵌入的模型，但实验缺少对模型鲁棒性的对比验证。Feng 等人[10]认为图的平滑性会导致对抗扰动在节点间传播，并针对这个问题通过添加一个对抗正则项，降低目标样本及其相连样本与预测值间的差异，结果表明添加了正则项的 GCN-GAD 对对抗扰动的敏感度下降，但实验中没有明确使用的对抗攻击方法。Wang 等人[11]提出忽略图的离散性而直接对邻接矩阵、特征矩阵加入扰动，针对随机丢弃边一种攻击方法，实验验证提出的 GraphDefense 方法能够保证对抗攻击后模型准确率提升 0.2 左右。图数据的离散性给 GCN 上的对抗训练带来了挑战，在邻接矩阵或特征矩阵上直接加入扰动的方法能够降低对抗训练方法的复杂度。

1.2　集成学习

集成学习被广泛研究用于提高模型的性能，通过结合多个单学习器，提升整体模型的泛化性。由于神经网络模型总会倾向于从数据集中提取类似的特征进行学习，对抗攻击在不同的图神经网络间也具有泛化性[12]。基于集成学习的对抗攻击防御方法，可以通过使不同子模型间具有不同的对抗子空间(adversarial space, Adv-SS)[13]，防止对抗攻击造成的影响在子模型间转移[14]。Kariyappa 等人[5]提出多样性训练降低子模型间损失函数的相关性。Pang 等人[4]提出了一种自适应的正则项，鼓励不同子模型的非极大预测值具有多样化。Yang 等人[15]基于数据中非鲁棒特征分布更普遍的发现，通过让子模型提取不同非鲁棒特征并集成模型学习能力，提升了模型在干净数据和攻击数据上的表现。上述基于集成学习的方法在图像领域取得了显著成果。

目前，将集成学习用于图领域以提升模型表现和模型鲁棒性的研究很少。张嘉杰等人[14]基于节点间的特征相似度重构了一个属性图，并分别基于结构信息和属性图进行预测，最后聚合二者的预测值作为返回结果。这种集成算法基于特征相似的节点以及相邻节点间通常具有相似标签的假设，对于属性信息进行了预处理，一定程度提升了模型表现，但在图结构受到扰动的情况下无法消除攻击影响，存在一定的局限性。Wu 等人[6]选取两个子模型分别从图的结构信息和属性信息进行学习，并在每轮迭代中平均两个子模型的置信度，将集成模型最有信心的预测值作为该节点的伪标签，并将该节点加入到训练集中，以此提升模型鲁棒性。该方法主要用于解决半监督学习下缺少标签的问题，没有考虑对抗攻击下图结构和节点属性上的变化。

1.3　非鲁棒特征研究

监督学习下，神经网络通过提取学习数据集中的特征提升模型能力，神经网络学习到的特征将直接决定其模型预测能力，Ilyas 等人[12]认为模型学习到的泛化性良好的特征是对抗攻击的基础，并通过在图像数据集上构建“鲁棒版数据集”和“非鲁棒版数据集”，证明了数据集中的非鲁棒特征会导致神经网络容易受到对抗攻击影响，因此非鲁棒特征在提升神经网络鲁棒性上具有研究价值。Yang 等人[15]通过模型卷积层后的嵌入向量提取图像数据中的非鲁棒特征，在提升模型鲁棒性的同时保证了干净数据集上的模型表现。

目前针对非鲁棒特征的研究大都集中在图像领域，但 Garg 等人[16]发现不受对抗攻击影响的鲁棒特征与图像数据的谱特征有关，说明图数据的拉普拉斯矩阵同样可能存在非鲁棒特征从而导致 GCN 的脆弱性。Jin 等人[17]的实验证明，去除对抗攻击的边和正常的边会对邻接矩阵的秩和奇异值产生不同的影响，说明对抗攻击生成过程所利用的特征具有特殊性，侧面印证了 GCN 学习到的特征中对于对抗攻击的易感性不同。由于 GCN 同时基于结构信息和节点属性信息进行模型训练，图数据上的非鲁棒特征研究应针对这两方面展开。文献[17]通过对比现实世界中的图和 metattack[18]攻击后的图，发现现实图中的相连节点大多倾向于拥有相似的属性特征，而对抗攻击会改变图的平滑度。文献[10]通过提升图的平滑度构造正则项提升了模型表现。上述研究说明图数据集中的非鲁棒特征可能与图的平滑度有关。

2　基于非鲁棒特征的集成对抗训练方法

受图像领域的非鲁棒特征提取方法启发，本文提出基于

非鲁棒特征的集成对抗训练方法 VDERG，考虑图数据的拓扑结构信息和节点属性信息，在模型的嵌入向量空间，分别通过与随机图的矩阵差异和特征平滑度差异获得梯度，对应在邻接矩阵和属性矩阵上进行迭代从而得到图数据中的非鲁棒特征。将非鲁棒特征作为对抗样本，让两个子模型分别在得到的结构非鲁棒特征和属性非鲁棒特征上进行对抗训练，最后对两个子模型的嵌入向量求和取平均，通过 softmax 函数得到节点预测标签。方法整体流程如图 1 所示。

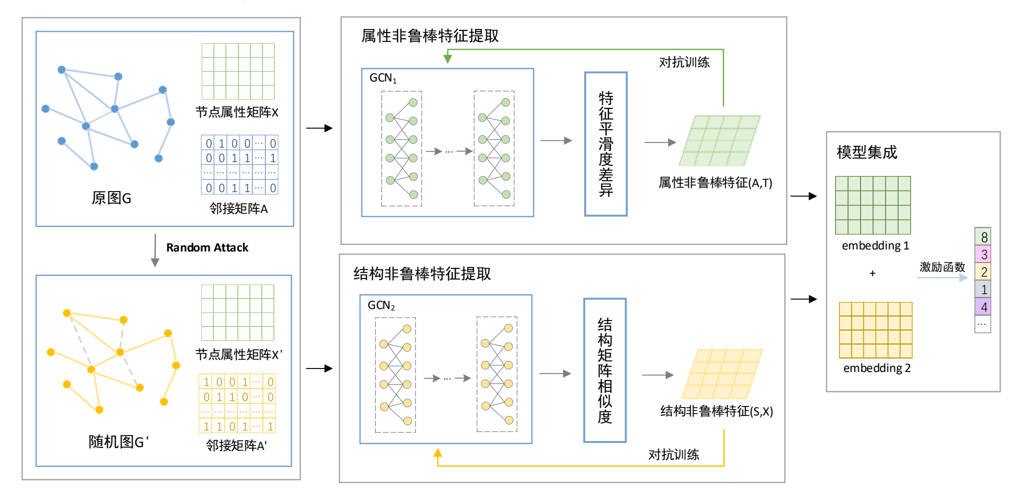


图 1　VDERG 方法流程图

Fig. 1　Flow chart of VDERG

2.1　问题描述

定义一个图为$\mathcal{G}=(\mathcal{V},\mathcal{E})$，其中$\mathcal{V}$为节点集合，包含 N 个节点$\{v_1,v_2,\cdots,v_N\}$，$\mathcal{E}$为边的集合。节点间的关系可以通过邻接矩阵$A\in\mathbb{R}^{N\times N}$进行表示，其中$A_{ij}$表示节点$v_i$和节点$v_j$间的关系。$X=[x_1,x_2,\ldots,x_N]\in\mathbb{R}^{N\times N}$表示节点特征矩阵，$x_i$表示节点$v_i$的特征向量。则一个图也可以用$\mathcal{G}=(A,X)$来表示。根据常见的节点分类任务设定，本文假定数据集中只有部分节点$\mathcal{V}_L=\{v_1,v_2,\cdots,v_L\}$带有标签$\mathcal{Y}_L=\{y_1,y_2,\cdots,y_L\}$，其中节点$v_i$的标签对应为$y_i$。则对于节点分类任务，给定图$\mathcal{G}=(A,X)$以及部分节点标签$\mathcal{Y}_L$，GCN 的目标是学习一个能够将节点映射到一组标签的函数$f_\theta:\mathcal{V}_L\to\mathcal{Y}_L$，并利用函数$f_\theta$对无标签的节点进行分类预测，学习过程可以由以下公式进行描述：

$$\min_{\theta}\mathcal{L}_{GCN}(\theta,A,X,\mathcal{Y}_L)=\sum_{v_i\in\mathcal{V}_L}\ell(f_\theta(X,A)_i,y_i) \tag{1}$$

其中，θ为需要学习的f_θ的参数，$f_\theta(X,A)_i$表示节点v_i的预测值，$\ell(\cdot,\cdot)$表示预测值和标签之间的差异，通常用交叉熵函数计算。目前最常用的 GCN 结构为两层 GCN[19]，即模型参数$\theta=(W_1,W_2)$，则函数f_θ可以进一步细化为

$$f_\theta(X,A)=\mathrm{softmax}\left(\hat{A}\sigma\left(\hat{A}XW_1\right)W_2\right) \tag{2}$$

其中，$\hat{A}=\tilde{D}^{-1/2}(A+I)\tilde{D}^{-1/2}$表示对邻接矩阵进行标准化，$\tilde{D}$表示$A+I$对角矩阵，$\tilde{D}_{ii}=1+\sum_j A_{ij}$；$\sigma$表示激励函数，常用 ReLU 函数。

基于上述定义，给定图$\mathcal{G}=(A,X)$和标签$\mathcal{Y}_L$，本文提出的 VDERG 算法将针对投毒攻击，在邻接矩阵A和特征矩阵X可能被投毒的前提下，学习 GCN 模型参数θ，通过对抗训练得到一个具有鲁棒性的 GCN 模型，提升对抗攻击下无标签节点上的预测分类表现。

2.2　非鲁棒特征提取

GCN 通过提取图数据中的特征学习节点的嵌入表示，提取过程中所利用的图数据特征中，一部分特征具有鲁棒性，即不易受到对抗攻击扰动的影响，反之则为非鲁棒特征，受到攻击后会使模型的表现下降。

设想最理想的非鲁棒特征提取情况：提取得到的扰动图中蕴涵所有可能干扰 GCN 的非鲁棒特征，且原图数据和扰动图数据间的差异巨大，但通过 GCN 模型后得到了相同的嵌入向量，如图 2 所示，则扰动图中包含的非鲁棒特征将对 GCN 生成节点嵌入表示产生致命影响。基于以上理论，本文对 GCN 在图数据上提取的非鲁棒特征作出如下定义：$\mathcal{G}=(A,X)$为原图数据的邻接矩阵及属性矩阵，$\mathcal{G}'=(A',X')$为随机生成的、与原图具有相同节点数但节点关系、属性特征不同的图数据。GCN 模型f的第l层从图$\mathcal{G}'$中提取出的、对应图$\mathcal{G}$的非鲁棒特征$\mathcal{R}=(S,T)$可以由下式定义：

$$\begin{aligned}&dis(\mathcal{G},\mathcal{G}')=\mathop{\arg\min}_{(S,T)}[\alpha L(f^l(S),f^l(A))+\beta R(f^l(T),f^l(X))],\\&\|S-A'\|_\infty\le\epsilon,\ \|T-X'\|_\infty\le\zeta\end{aligned} \tag{3}$$

其中，$f^l(\cdot)$表示 GCN 模型第l个隐藏层在激励函数(如 ReLU)前的输出。考虑到对抗攻击可以通过修改节点间关系或节点属性对 GCN 模型产生干扰，式(3)代表的特征提取过程分别从邻接矩阵和属性矩阵两方面进行约束优化，目标是从图$\mathcal{G}'$中提取出可能让 GCN 模型混淆识别为图$\mathcal{G}$的特征，即图$\mathcal{G}$中的非鲁棒特征。

式(3)的第一项$L(f^l(S),f^l(A))$通过最小化原图邻接矩阵A和提取邻接特征S在嵌入空间的差异，使从节点关系中提取的特征接近 GCN 学习到的节点关系信息。可以通过约束$f^l(S)$和$f^l(A)$的 F 范数实现上述目标，即可以将第一项重写为

$$L(f^l(S),f^l(A))=\|f^l(S)-f^l(A)\|_F^2 \tag{4}$$

而第二项$R(f^l(T),f^l(X))$则考虑从节点的属性信息中提取特征。对抗攻击在连接属性差异大的节点或删除相似节点间链接时，会降低图的平滑度，因此本文考虑通过最小化原图属性矩阵X和提取属性特征T间的特征平滑度差异，使从节点属性中提取的特征接近 GCN 学习的节点属性信息，即式(3)的第二项可以被重写为

$$\begin{aligned}R(f^l(T),f^l(X))=\frac{1}{2}\|&\sum_{n,m=1}^{N}S_{nm}(f^l(t_n)-f^l(t_m))^2-\\&\sum_{n,m=1}^{N}A_{nm}(f^l(x_n)-f^l(x_m))^2\|\end{aligned} \tag{5}$$

其中，A表示图数据的邻接矩阵，A_{nm}表示节点v_n和v_m相连，$(x_n-x_n)^2$衡量了节点v_n和v_m通过 GCN 模型得到的嵌入向量间的差异。$\frac{1}{2}\sum_{n,m=1}^{N}A_{nm}(x_n-x_m)^2$衡量了图$(A,X)$的特征平滑度差异，$\frac{1}{2}\sum_{n,m=1}^{N}S_{nm}(t_n-t_m)^2$同理。通过约束特征平滑度差异进行属性特征提取，充分考虑了现实中攻击者常常将差异较大的节点相连以降低模型预测能力的攻击特点。

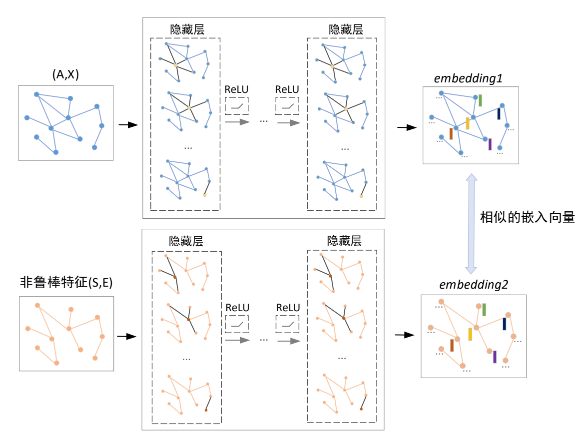


图 2　非鲁棒特征示意图

Fig. 2　An illustrative example on non-robust features

2.3　基于非鲁棒特征的集成对抗训练

集成学习作为一种训练思路能够用于提升模型鲁棒性，通过让集成学习中的子模型分别学习到不同的特征，能够在保持简单模型结构的前提下提升模型表现。而若能让不同子模型学习到不同的非鲁棒特征，则能避免对抗攻击的泛化性影响到所有子模型，提升集成后的模型表现。基于上述理论，本文采用集成学习思想，采用两个子模型分别从节点关系和节点属性两个方面提取图数据中的非鲁棒特征，并利用提取的非鲁棒特征对模型进行对抗训练。对抗训练常通过在样本中加入微小扰动，使神经网络适应扰动从而提升对抗样本上的模型鲁棒性。但图数据作为非欧几里德数据结构，无法通过梯度相关方法构造对抗样本。因此，通过提取的特征对模型进行对抗训练，避开了构造对抗样本过程中需要考虑的数据离散问题，更加简单且具有可解释性。

2.3.1 基于节点关系的非鲁棒特征学习方法

参考式(1)和(2)，第一个子模型从邻接矩阵中提取节点关系所含非鲁棒特征(S,X)的过程可以用下式表示：

$$\arg\min_{S}\mathcal{L}_a=\|f_1{}^2(S)-f_1{}^2(A)\|_F^2,\quad \|S-A'\|_\infty\le\epsilon \tag{6}$$

其中，$f_1{}^2$表示第一个子模型的第二层卷积层后、激励函数前的嵌入向量。通过约束特征S和随机图邻接矩阵A'间的差异小于ϵ，最小化S和原图邻接矩阵A在嵌入空间中的差异距离，从随机图$\mathcal{G}'$的邻接矩阵中提取出与随机图$\mathcal{G}'$相似，但会错使 GCN 模型预测为图$\mathcal{G}$的非鲁棒特征。则第一个子模型进行对抗训练时的目标损失函数为

$$\arg\min_{S}\mathcal{L}_1=\mathcal{L}_{f_1}=\mathcal{L}_{GCN}(\theta_1,S,X,\mathcal{Y}_L),\quad \|S-A'\|_\infty\le\epsilon \tag{7}$$

其中，$\mathcal{L}_{GCN}(\theta_1,S,X,\mathcal{Y}_L)$是第一个 GCN 子模型在输入特征$(S,X)$上的损失函数。通过最小化式(7)，能够训练第一个模型学习到节点关系中包含的非鲁棒特征，提升模型的鲁棒性。

2.3.2 基于节点属性的非鲁棒特征学习方法

参考式(1)和(3)，同样基于 GCN 第二层卷积层后的嵌入向量，第二个子模型从属性矩阵中提取节点属性所含非鲁棒特征(A,T)的过程可以用下式表示：

$$\arg\min_{T}\mathcal{L}_f=\frac{1}{2}\|\sum_{n,m=1}^{N}S_{nm}(f_2{}^2(t_n)-f_2{}^2(t_m))^2-\sum_{n,m=1}^{N}A_{nm}(f_2{}^2(x_n)-f_2{}^2(x_m))^2\|,\quad \|T-X'\|_\infty\le\zeta \tag{8}$$

其中，$f_2{}^2$表示第二个子模型的第二层卷积层后、激励函数前的嵌入向量。类似地，约束特征T和随机图属性矩阵的差异小于ζ，并最小化T和原图属性矩阵X在嵌入空间中的特征平滑度差异，以此从随机图$\mathcal{G}'$的属性矩阵中提取出对应图$\mathcal{G}$的非鲁棒特征。则第二个子模型进行对抗训练时的目标损失函数为

$$\arg\min_{T}\mathcal{L}_2=\mathcal{L}_{f_2}=\mathcal{L}_{GCN}(\theta_2,A,T,\mathcal{Y}_L),\quad \|T-X'\|_\infty\le\zeta \tag{9}$$

其中，$\mathcal{L}_{GCN}(\theta_2,A,T,\mathcal{Y}_L)$是第二个 GCN 子模型在输入特征$(A,T)$上的损失函数。式(9)能够训练第二个子模型从节点属性的角度学习非鲁棒特征，降低对抗攻击的影响效果。

2.3.3 基于集成学习的对抗训练策略

基于上述非鲁棒特征学习方法，本文提出的 VDERG 的对抗训练过程如下：首先随机生成两个 GCN 子模型，在每轮迭代中根据输入图生成节点数相同的随机图，通过随机梯度下降优化式(6)和(8)，借助随机图分别从邻接矩阵和属性矩阵中提取输入图的非鲁棒特征。然后基于节点关系和节点属性中的非鲁棒特征分别对两个子模型进行对抗训练，利用式(7)和(9)的交叉熵损失函数优化子模型参数，通过 Adam 对网络参数进行优化，最后对两个子模型得到的节点嵌入向量求和取均值，经过 softmax 函数得到最终模型的预测结果。伪代码如算法 1 所示。

算法 1　VDERG 的训练策略

输入：邻接矩阵A，属性矩阵X，标签$\mathcal{Y}_L$，特征提取过程轮数N_1、N_2，步长α、β，学习率η。

输出：集成 GCN 模型的参数θ_1、θ_2，节点预测结果。

```
Randomly initialize θ1, θ2  //初始化 2 个 GCN 子模型的参数;
for e in range(E):
    (A′,X′)←random_attack((A,X),ptb_rate=1)  /*通过在输入图数据上实施扰动率 1.0 的随机攻击生成随机图*/
    Initialize S←A′  //利用随机图的邻接矩阵初始化特征S
    for i in N1:
        通过式(6)以步长α更新S得到(S,X)  //基于嵌入空间从邻接矩阵提取非鲁棒特征
    end
    g1←∂L_GCN(θ1,S,X,Y_L)/∂θ1
    θ1←θ1−ηg1  //更新第一个子模型参数
    Initialize T←X′  //利用随机图的属性矩阵初始化特征T
    for j in N2:
        通过式(8)以步长β更新T得到(A,T)  /*基于嵌入空间的特征平滑度差异提取非鲁棒特征*/
    end
    g2←∂L_GCN(θ2,A,T,Y_L)/∂θ2
    θ2←θ2−ηg2  //更新第二个子模型参数
    embedding1=g1(S,X)  //基于结构非鲁棒特征得到嵌入向量
    embedding2=g2(A,T)  //基于属性非鲁棒特征得到嵌入向量
    pred=softmax((embedding1+embedding2)/2)
end
```

3　实验结果与分析

3.1　数据集描述

本文选取了图领域常见的三种引文网络数据集作为数据集进行节点分类任务实验，数据集的详细信息如表 1 所示。实验中，参考著名攻击算法 Metattack 的数据集划分方法，本文将所有数据集按照 10%和 90%的比例随机分割为有标签集和无标签集，再进一步把有标签集按照 50%和 50%的比例分

为训练集和验证集。

表 1　数据集描述

Tab. 1　Data description

数据集	节点数	特征	类别	边
Cora	2708	1433	7	5278
Citeseer	3327	3703	6	4552
PubMed	19717	500	3	44324

3.2　模型效果对比

为了验证本文所提出的 VDERG 的对抗攻击防御能力，本文基于对抗攻击算法 Metattack，将 VDERG 与目前效果最好的几种 GCN 防御算法在节点分类准确率上进行对比评估。Metattack 有五个变种，在数据集 Cora 和 Citeseer 上本文采用攻击效果最好的 Meta-Self 变种进行实验；在数据集 Pubmed 数据集上，出于节省时间的内存的考虑，本文采用和 Meta-Self 变种相似的 A-Meta-Self 变种进行实验。实验针对从 0 到 20%的扰动率进行了实验，每次提升 5%扰动率，参考实验结果如表 2 所示，其中 GCN、GAT、GCN-Jaccard、Pro-GNN 的实验结果来自文献[17]，SimP-GCN 的实验结果来自原论文。为了使模型结果更加客观、消除深度学习训练过程中的随机性，所有实验均重复了 10 次。

表 2　全局攻击(metattack)下节点分类任务表现对比

Tab. 2　Node classification performance under non-targeted attack (metattack)

数据集	扰动率/%	GCN[19]	GAT[20]	Pro-GNN[17]	SimP-GCN [21]	VDERG
Cora	0	83.50±0.44	83.97±0.65	83.42±0.52	81.81±0.62	84.26±0.43
	5	76.55±0.79	80.44±0.74	82.78±0.39	76.43±1.98	83.98±0.63
	10	70.39±1.28	75.61±0.59	79.03±0.59	73.27±1.93	82.72±1.38
	15	65.10±0.71	65.10±0.71	76.40±1.27	70.75±3.98	81.70±0.71
	20	59.56±2.72	59.94±0.92	73.32±1.56	66.63±6.87	80.23±1.21
Citeseer	0	71.96±0.55	73.26±0.83	73.28±0.69	73.76.±0.78	75.01±1.09
	5	70.88±0.62	72.89±0.83	73.09±0.34	73.12±0.85	74.16±0.66
	10	67.55±0.89	70.63±0.48	72.51±0.75	72.38±0.67	73.76±0.38
	15	64.52±1.11	69.02±1.09	72.03±1.11	71.75±1.54	73.52±0.81
	20	62.03±3.49	61.04±1.52	70.02±2.28	69.37±1.50	73.41±1.23
PubMed	0	87.19±0.09	83.73±0.40	87.33±0.18	87.59.±0.10	87.91±0.23
	5	83.09±0.13	78.00±0.44	87.25±0.09	86.79±0.12	87.87±0.14
	10	81.21±0.09	74.93±0.38	87.25±0.09	86.01±0.10	87.76±0.13
	15	78.66±0.12	71.13±0.51	87.20±0.09	85.49±0.11	87.55±0.11
	20	77.35±0.19	68.21±0.96	87.15±0.15	85.37±0.12	87.41±0.15

从表 2 的结果可以看出，在扰动率为 0 时，VDERG 在 Cora、Citeseer 和 PubMed 数据集上的模型准确率分别在目前最优模型的基础上提升了 0.84%、1.25%和 0.32%，说明 VDERGE 通过集成节点属性特征和结构特征能够更全面地学习到图数据蕴涵的信息，并且通过非鲁棒特征进行对抗训练不仅能够提升模型鲁棒性，还能提升模型在干净数据集上的表现。

针对扰动率为 5%至 20%的情况，VDERG 在三个数据集上都比现有最优模型取得了更高的准确率，扰动率的提升并没有使 VDERG 像原始 GCN 那样在准确率上产生明显的下降。在扰动率提升的过程中，相较其他方法，VDERG 的模型准确率下降更为缓慢，表现出了更强的鲁棒性。同现其他分类器相比，VDERG 在 Cora 数据集上的表现提升最为明显，当扰动率为 20%时，VDERG 的准确率比现有最优模型高了 6.91%。

3.3　集成与单一特征学习对比

为了研究集成方法在替身模型表现过程中的有效性，本小节对比在集成过程中仅考虑结构信息或属性信息的模型表现，实验结果如表 3 所示，图中分别展示了只从结构信息提取非鲁棒特征的 VDERG-structure 和只从属性信息提取非鲁棒特征的 VDERG-features 在数据集 Cora 和 Citeseer 上的结果。从表中可以看出，虽然在 Cora 数据集的原始数据上 VDERG 的效果略逊于单独考虑结构信息，但在 Citeseer 数据集上以及受到对抗攻击时，综合考虑了结构信息和属性信息的 VDERG 都取得了最好的分类效果，说明本文提出的集成策略能够有效提升模型鲁棒性，提高对抗攻击下的图信息表征能力。同时，实验结果表明，仅仅基于结构信息中的非鲁棒特征进行对抗训练比仅基于属性信息的方法效果更好，这是因为基于特征平滑度差异提取非鲁棒特征时可能造成孤立节点的过平滑，而 VDERG 综合考虑结构信息能够有效弥补这一缺陷。

表 3　结构和属性消融实验对比

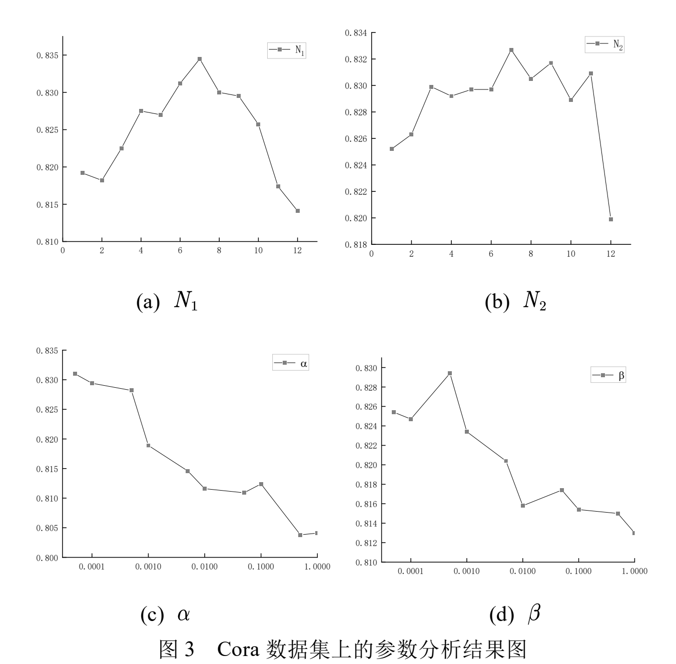
Tab. 3　The comparison of structure and features ablation

数据集	扰动率/%	VDERG-structure	VDERG-features	VDERG
Cora	0	85.51±0.30	84.48±0.53	84.26±0.43
	5	83.82±0.75	83.65±1.21	83.98±0.63
	10	82.10±0.85	81.50±1.40	82.72±1.38
	15	81.69±1.36	81.37±1.18	81.70±0.71
	20	80.19±1.31	80.12±1.02	80.23±1.21
Citeseer	0	74.46±1.06	74.75±0.71	75.01±1.09
	5	73.31±1.04	73.69±1.84	74.16±0.66
	10	73.02±0.62	73.02±0.50	73.76±0.38
	15	73.34±0.47	72.71±1.57	73.52±0.81
	20	72.18±0.80	72.86±0.95	73.41±1.23

3.4　模型不同参数对比

对于本文提出的 VDERG 策略，非鲁棒特征的提取效率至关重要。因此，本小节在 10%扰动率的 Metattack 攻击下的 Cora 数据集上，研究分析了结构非鲁棒特征提取过程中的步长α和轮数N_1，以及属性非鲁棒特征提取过程中的步长β和轮数N_2对 VDERG 效果的影响，实验结果如图 3 所示。本文设定步长α和步长β的变化范围为 5e-5 到 1，轮数N_1和N_2的变化范围为 1 到 12。从图 3 中可以看出，在两种非鲁棒特征提取过程中，模型的性能随着迭代轮数的变化呈现先上升后下降的趋势，对于结构信息的特征提取，最佳轮数为 7；在迭代轮数为 8 至 11 时，属性信息特征提取随迭代轮数变化的曲线波动更明显，同样在轮数为 7 时取得最好模型效果，迭代轮数达到 11 后模型性能开始显著下降。此外，图中还可

以看出模型在两种非鲁棒特征提取过程中随步长的改变有着相似的变化趋势，都呈现先上升后下降的形势，结构非鲁棒特征提取的最佳迭代步长为 5e-5 ，属性非鲁棒特征提取的最佳迭代步长为 5e-4。



(a) N_1　(b) N_2

(c) α　(d) β

图 3　Cora 数据集上的参数分析结果图

Fig. 3　Results of parameter analysis on Cora dataset

4　结束语

本文针对图卷积神经网络提出了一种基于非鲁棒特征的集成对抗训练策略。通过从图卷积层后的嵌入向量中提取非鲁棒特征进行对抗训练，能够绕开直接构造对抗样本时面临的数据离散等问题。为了充分利用图数据信息，本文提出的策略分别从拓扑结构和节点属性两方面的信息出发，借助随机图分别提取输入图中蕴涵的非鲁棒特征，利用非鲁棒特征对两个子模型分别进行对抗训练，并最终集成两个子模型得到的嵌入向量，得到节点预测分类。在引文网络上的实验证明，在 Cora、Citeseer、PubMed 原始数据集上，该策略较目前最优模型分别提升了 0.84%、1.25%和 0.32%的准确率；Cora 数据集上，面对 20%扰动率的对抗攻击时，能够比现有最优模型提升 6.91%的准确率，以上实验结果充分证明了本文提出的策略能够提升模型在干净数据和攻击图上的节点分类任务表现。

通过对比集成模型与单一特征学习模型的实验结果可以看出，不论是在原始数据集上还是在攻击情景下，集成结构拓扑和节点属性的策略都比单从一个方面进行对抗训练的模型效果更好。

在接下来的工作中，本文计划针对包含较多孤立节点的数据集提升非鲁棒特征的提取效果，研究其他图神经网络模型结构对非鲁棒特征的敏感度以及学习表现，更深入探索图数据中的非鲁棒特征与对抗攻击间的关系。

参考文献：

[1] Zügner D, Akbarnejad A, Günnemann S. Adversarial attacks on neural networks for graph data [C]// Proc of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2018: 2847-2856.

[2] Dai Hanjun, Li Hui, Tian Tian, *et al.* Adversarial attack on graph structured data [C]// Proc of the 35th International Conference on Machine Learning. [S. l.] : PMLR Press, 2018: 1115-1124.

[3] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [C]// Proc of the 6th ICLR, 2015.

[4] Pang Tianyu, Xu Kun, Du Chao, *et al.* Improving adversarial robustness via promoting ensemble diversity [C]// Proc of the 36th International Conference on Machine Learning. [S. l.] : PMLR Press, 2019: 4970-4979.

[5] Kariyappa S, Qureshi M K. Improving adversarial robustness of ensembles with diversity training [EB/OL]. (2019-01-28) [2022-01-04]. https://arxiv. org/pdf/1901. 09981.

[6] Wu Xuguang, Wu Huijun, Zhou Xu, *et al.* CoG: a two-view co-training framework for defending adversarial attacks on graph [EB/OL]. (2021-9-12) [2022-01-04]. https://arxiv. org/pdf/2109. 05558.

[7] Sun Lichao, Dou Yingtong, Yang C, *et al.* Adversarial attack and defense on graph data: A survey [EB/OL]. (2020-07-12) [2022-01-04]. https://arxiv. org/pdf/1812. 10528.

[8] Dai Quanyu, Shen Xiao, Zhang Liang, *et al.* Adversarial training methods for network embedding [C]// Proc of WWW Conference. New York: ACM Press, 2019: 329-339.

[9] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations [C]// Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014: 701-710.

[10] Feng Fuli, He Xiangnan, Tang Jie, *et al.* Graph adversarial training: Dynamically regularizing based on graph structure [J]. IEEE Trans on Knowledge and Data Engineering, 2019, 33 (6): 2493-2504.

[11] Wang Xiaoyun, Liu Xuanqing, Hsieh C J. GraphDefense: Towards robust graph convolutional networks [EB/OL]. (2019-11-11) [2022-01-04]. https://arxiv. org/pdf/1911. 04429.

[12] Ilyas A, Santurkar S, Tsipras D, *et al.* Adversarial examples are not bugs, they are features [J]. Advances in Neural Information Processing Systems, 2019, 32: 125-136.

[13] Tramèr F, Papernot N, Goodfellow I, *et al.* The space of transferable adversarial examples [EB/OL]. (2017-05-23) [2022-01-04]. https://arxiv. org/pdf/1704. 03453.

[14] 张嘉杰，过弋，王家辉，等. 基于特征和结构信息增强的图神经网络集成学习框架 [J/OL]. 计算机应用研究, 2021, 39 (3) . (2021-12-07) [2022-01-04]. https://www. arocmag. com/article/02-2022-03-033. html. (Zhang Jiajie, Guo Yi, Wang Jiahui, *et al.* Ensemble learning framework for graph neural netword with feature and structure enhancement [J/OL]. Application Reasearch of Computers, 2021, 39 (3) . (2021-12-07) [2022-01-04]. https://www. arocmag. com/article/02-2022-03-033. html.)

[15] Yang Huanrui, Zhang Jingyang, Dong Hongliang, *et al.* DVERGE: diversifying vulnerabilities for enhanced robust generation of ensembles [J]. Advances in Neural Information Processing Systems, 2020, 33: 5505-5515.

[16] Garg S, Sharan V, Zhang B H, *et al.* A spectral view of adversarially robust features [J]. Advances in Neural Information Processing Systems, 2018, 31: 10159–10169.

[17] Jin Wei, Ma Yao, Liu Xiaorui, *et al.* Graph structure learning for robust graph neural networks [C]// Proc of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2020: 66-74.

[18] Zügner D, Günnemann S. Adversarial attacks on graph neural networks via meta learning [EB/OL]. (2019-02-22) [2022-01-04]. https://arxiv. org/pdf/1902. 08412.

[19] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [EB/OL]. (2017-02-22) [2022-01-04]. https://arxiv. org/pdf/1609. 02907.

[20] Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks

[C]// Proc of the 6th ICLR, 2018.

[21] Jin Wei, Derr T, Wang Yiqi, *et al*. Node similarity preserving graph convolutional networks [C]// Proc of the 14th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2021: 148-156.